Lineare Regression (IQ)

Vorhersage des IQ mit linearer Regression

In dieser Aufgabe soll mittels linearer Regression der Intelligenz-Quotient (IQ) von Kindern vorhergesagt werden.

Daten importieren und anschauen

Laden Sie die Daten unter https://oc.informatik.hs-mannheim.de/s/K2nJQngd8N6o3M5/download in einen Data Frame. Die Daten sind im RDS-Format gespeichert. Der Datensatz stammt von Martin Brand (brandt@psychologie.uni-mannheim.de, https://www.sowi.uni-mannheim.de/kuhlmann/team/akademische-mitarbeiterinnen-und-mitarbeiter/brandt-martin/). Die Spalten bedeuten:

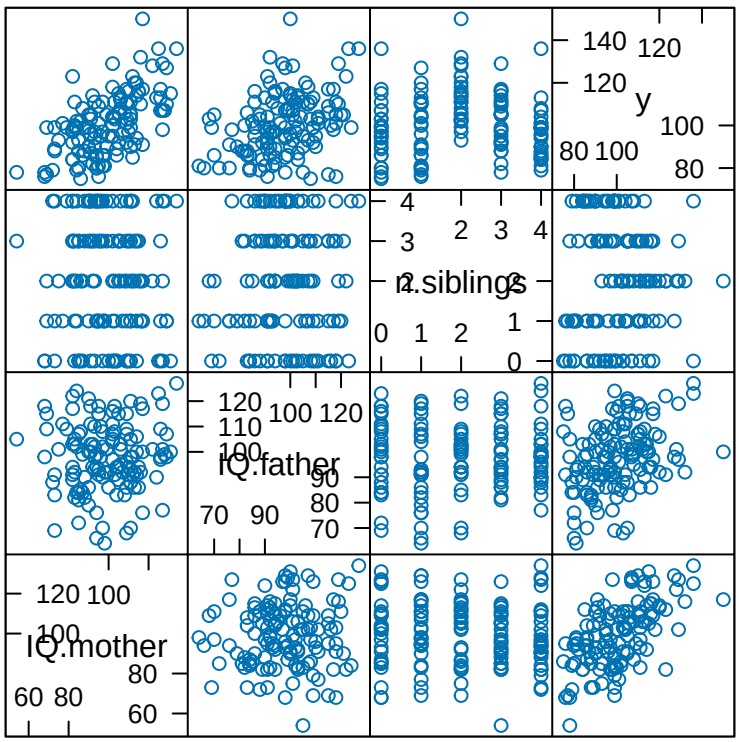
- `id`: Eindeutige Nummer
- `IQ.mother`: IQ der Mutter
- `IQ.father`: IQ des Vaters
- `n.siblings`: Anzahl der Geschwister
- `IQ.child`: IQ des Kindes

```
data = readRDS(url("https://oc.informatik.hs-mannheim.de/s/K2nJQngd8N6o3M5/download", "rb"))
```

Verschaffen Sie sich einen ersten Überblick über die Daten. Plotten Sie dazu ein Mehrfach-Diagramm, das Scatterplots aller vier Variablen (drei Merkmale und Response `IQ.child`) untereinander zeigt. D.h. jede Variable soll mit jeder anderen paarweise verglichen werden.

Tipp: Mit der `caret::featurePlot`-Funktion lässt sich das gut erreichen.

```
featurePlot(x = data[, c("IQ.mother", "IQ.father", "n.siblings")],
            y = data$IQ.child,
            plot = "pairs",
            auto.key = list(columns=3))
```



Scatter Plot Matrix

Lineare Regression

Wenden Sie für diesen Datensatz eine lineare Regression an, um den IQ von Kindern in Abhängigkeit der Merkmale schätzen zu können. Die lineare Regression soll im Objekt `lm.iq` gespeichert werden.

```
# erkläre IQ des Kindes durch IQ von Mutter, Vater und Anzahl der Geschwister
lm.iq = lm(IQ.child ~ IQ.mother + IQ.father + n.siblings, data = data)
```

Merkmale

Welche Merkmale spielen eine Rolle? Interpretieren Sie das Ergebnis.

```
summary(lm.iq)
```

```
##
## Call:
## lm(formula = IQ.child ~ IQ.mother + IQ.father + n.siblings, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.404  -7.985  -0.163   6.568  39.073
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.96113    8.99908   1.329    0.186
## IQ.mother    0.53812    0.05906   9.111 1.58e-15 ***
## IQ.father    0.35134    0.07069   4.970 2.13e-06 ***
## n.siblings   0.43556    0.65779   0.662    0.509
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 10.73 on 126 degrees of freedom
## Multiple R-squared:  0.4713, Adjusted R-squared:  0.4587
## F-statistic: 37.44 on 3 and 126 DF,  p-value: < 2.2e-16
```

Residuals: Zusammenfassung der Differenzen zwischen vorhergesagten und tatsächlichen Werten

Koeffizienten: Intercept: IQ des Kindes, wenn alle Prädiktoren = 0 Prädiktoren: für jede Einheitserhöhung der Variable wird abhängige Variable um x Einheiten erhöht (z. B. 0.53812)

Signif. codes: Wie signifikant sind die einzelnen Koeffizienten? 0 also *** bedeutet hoch signifikant, kein Symbol bedeutet nicht signifikant

Residual Standard Error: Durchschnittliche Abweichung der beobachteten Werte von vorhergesagten

Multiple R-squared: Anteil der Variation in IQ.child, der durch unabhängige Variablen erklärt wird

Adjusted R-squared: berücksichtigt Anzahl der verwendeten Prädikatoren (R-squared kann dazu neigen, zu steigen, je mehr Prädiktoren zum Modell hinzugefügt werden)

F-statistic: explained variance/unexplained variance -> bewertet statistische Signifikanz der Regression

Interpretation: Betrachte die Signifikanz der Koeffizienten. Der IQ der Mutter und der IQ des Vaters sind mit *** gekennzeichnet, was darauf hinweist, dass diese Werte mit dem IQ des Kindes korrelieren. Bei Betrachtung des t- und Pr-Wertes, ist erkennbar, dass t bei der Mutter höher ist als beim Vater und somit auch der Wert von Pr(>|t|) geringer. Das bedeutet, dass der IQ des Kindes mehr von dem IQ der Mutter abhängt, als von dem des Vaters.

Die Anzahl der Geschwister ist allerdings nicht signifikant für den IQ des Kindes, da der t-Wert sehr klein ist. Insgesamt werden etwa 47,13% der Variation im IQ des Kindes durch die im Modell enthaltenen unabhängigen Variablen erklärt.

Der hohe F-Wert zusammen mit dem kleinen p-Wert deuten darauf hin, dass das gesamte Modell statistisch signifikant ist.

RSE

Berechnen Sie den RSE. Was gibt dieser an?

```
rse = summary(lm.iq)$sigma
cat("Residual Standard Error (RSE):", round(rse, 2))
```

```
## Residual Standard Error (RSE): 10.73
```

RSE gibt die durchschnittliche Abweichung der beobachteten Werte von den vorhergesagten Werten an. Das bedeutet, der tatsächliche IQ des Kindes weicht von den vorhergesagten Werten durchschnittlich um 10,73 Einheiten ab.

Anteil erklärter Varianz

Wie groß (in Prozent) ist der Anteil der erklärten Varianz? Ist das Ergebnis zufriedenstellend?

```
r_squared = summary(lm.iq)$r.squared
cat("Multiple R-squared (R^2):", round(r_squared, 4)*100, "%")
```

```
## Multiple R-squared (R^2): 47.13 %
```

Insgesamt werden etwa 47,13% der Variation im IQ des Kindes durch die im Modell enthaltenen unabhängigen Variablen erklärt. Es ist schwierig eine genaue Aussage darüber zu treffen, ob das Ergebnis "zufriedenstellend" ist. Es wurden Merkmale untersucht, die definitiv mit dem IQ des Kindes korrelieren. Allerdings lässt sich der IQ des Kindes auch nicht ausschließlich durch den IQ der Eltern vorhersagen.

Geschwister-Anzahl als Faktor

Anwendung

Wandeln Sie den Datensatz so um, dass die Anzahl der Geschwister (`n.siblings`) ein Faktor ist und wenden Sie darauf die lineare Regression an.

```
data$n.siblings = as.factor(data$n.siblings)
lm.iq_factor = lm(IQ.child ~ IQ.mother + IQ.father + n.siblings, data = data)
summary(lm.iq_factor)
```

```
##
## Call:
## lm(formula = IQ.child ~ IQ.mother + IQ.father + n.siblings, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.7219  -5.4485  -0.3552   4.1449  27.3896
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.4584     7.4383   1.137  0.25769
## IQ.mother     0.5361     0.0478  11.214  < 2e-16 ***
## IQ.father     0.3481     0.0578   6.022 1.84e-08 ***
## n.siblings1   1.7058     2.4251   0.703  0.48315
## n.siblings2  17.3342     2.4525   7.068 1.03e-10 ***
## n.siblings3   6.7146     2.4049   2.792  0.00607 **
## n.siblings4   0.1098     2.3593   0.047  0.96296
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.659 on 123 degrees of freedom
## Multiple R-squared:  0.6637, Adjusted R-squared:  0.6473
## F-statistic: 40.46 on 6 and 123 DF,  p-value: < 2.2e-16
```

R^2 hat sich verändert. Interpretieren bzw. begründen Sie das Ergebnis:

Der Wert beträgt nun 66,37% im Vergleich zu 47,13% bei der vorherigen Auswertung. Es ist auffällig, dass die Anzahl der Geschwister nun doch mit dem IQ des Kindes korreliert, vor allem bei einer Anzahl von zwei oder drei Geschwistern im Gegensatz zu 0 Geschwistern. Bei einem oder vier Geschwistern gibt es allerdings keinen signifikanten Zusammenhang mit dem IQ.

R^2 ist höher, da nun diese unterschiedliche Auswirkungen für die unterschiedlichen Geschwisterkategorien berücksichtigt werden können. Es gibt insegesamt mehr Koeffizienten, und insgesamt auch mehr Koeffizienten, die mit dem IQ korrelieren, womit sich die Erhöhung des R^2 Wertes begründen lässt.

Schätzung für Testdaten

Abschließend soll mittels Cross-Validation überprüft werden, wie gut dieses Verfahren für unbekannte Testdaten funktioniert. Überlegen Sie sich ein passendes Verfahren und bestimmen Sie damit die R^2-Statistik.

```
# use k-Fold-Cross-Validation
train_control = trainControl(method = "cv", number = 10)

cv_res = train(IQ.child ~ IQ.mother + IQ.father + n.siblings,
                    data = data,
                    method = "lm",
                    preProcess = c("center", "scale"), # Daten werden zentriert und skaliert
```

```
                    trControl = train_control)

print(cv_res)
```

```
## Linear Regression
##
## 130 samples
##   3 predictor
##
## Pre-processing: centered (6), scaled (6)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 117, 116, 117, 116, 116, 117, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   8.707612  0.6789565  6.811998
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Cross Validation erstellt mehrere Trainings- und Testsets und überprüft somit, wie gut das Verfahren für unbekannte Testdaten funktioniert. R^2 beträgt hier ca. 67%. das Verfahren kann also auch bei unbekannten Daten ca. 67% der Variation im IQ der Kinder erklären.